but an accumulation of plain, unadorned facts available to any one's inspection, it seems useless to try to bolster either of them up by the dialectic methods of a lawyer's appeal to the jury.

RAYMOND PEARL.

June 21, 1911.

ON THE FORMATION OF CORRELATION AND CONTINGENCY TABLES WHEN THE NUMBER OF COMBINATIONS IS LARGE

IN earlier numbers of this JOURNAL two papers on that useful tool, the correlation coefficient, have appeared. The first[1] explains and illustrates a convenient method of carrying out the arithmetical routine of calculation, while the second, by Professor Jennings,[2] describes a method for obtaining the coefficient for symmetrical tables without the labor of actually rendering the tables themselves symmetrical.

The purpose of this note is to point out a method of preparing correlation tables where the number of combinations is large. Such tables are not infrequently needed. Suppose, for

[1] Harris, J. Arthur, "The Arithmetic of the Product Moment Method of Calculating the Coefficient of Correlation," AMER. NAT., Vol. 44, pp. 693–699, 1910.

In a note on this method of calculating the coefficient of correlation, Professor Jennings (AMER. NAT., Vol. 45, p. 413, 1911) suggests reduction in size of the moments by designating the lowest grade by 0 and the succeeding ones by 1, 2, 3, ⋯ n. In this he is quite justified. I have frequently used the scheme he suggests during the last several years, but I did not refer to it particularly in my note, and for two reasons. First, I thought the point sufficiently covered by the statement that the rough moments may be taken about any arbitrary point as origin, and by the suggestion that when the range is very great it may pay to use the conventional methods in calculating the standard deviations. Second, according to my experience it is better, whenever possible, to keep the actual values. When one uses a mechanical calculator the arithmetical routine is (after a little practise) not out of proportion to the advantages. Under many circumstances these are very great: (a) all the values have a direct biological (physical) significance, (b) the means of arrays may at once be obtained for testing linearity of regression, (c) tables for different lots of material may be combined or separated at will by merely summing or segregating their moments, and finally (d) I shall show in a forthcoming paper how these moments, once calculated, may be of much service in obtaining some of the more difficult correlations.

[2] Jennings, H. S., "Computing Correlation in Cases where Symmetrical Tables are Commonly Used," AMER. NAT., Vol. 45, pp. 123–128, 1910.

instance, that one wishes to correlate between the hatching quality of the eggs of sisters in the domestic fowl, as Pearl and Surface[3] have actually done. If each family be composed of only ten pullets and there be only fifty families the number of entries in the symmetrical correlation table will be $10 \times 9 \times 50 = 4{,}500$. Or again, if one be interested in determining whether the dimensions or proportions of the blood corpuscles differ from individual to individual in an animal, say the tadpole,[4] and have measurements of twenty-five corpuscles in each of 100 individuals, he may have to form a correlation table of 60,000 entries. Much larger tables than this have been formed. The labor is of course excessive, and this has been one of the factors limiting their application to problems of morphology, physiology and heredity.

In many cases the routine, as I have found from considerable experience, can be profitably carried out as follows.

The individuals of each class are seriated separately and the frequencies entered in horizontal rows in a table of vertical columns, each devoted to one of the grades of variates, g_{m+1}, g_{m+2}, $\ldots g_{m+n}$. A second table, exactly like the first in width of column and row is prepared and cut into strips by columns. Each of these columns is moved successively across the surface of the original table, and the frequencies which are in juxtaposition are multiplied together and their products summed and entered on a correlation blank, in the compartment corresponding to the captions of the two columns. This is repeated for all the columns *except the one identical with the strip*. If the strip be for grade g_{m+2}, the multiplications and summations from once passing it over the original table give the whole relative array associated with it as subject, except the frequencies of the diagonal cell, $g_{m+2} - g_{m+2}$.[5] To obtain these each frequency on a strip is multiplied by itself less one and the products summed.

It is not absolutely necessary, since the table is symmetrical,

[3] Pearl, R., and Surface, F. M., "Data on Certain Factors Influencing the Fertility and Hatching of Eggs," Bull. Me. Ag. Exp. Sta., No. 168, pp. 147–151, 1909.

[4] For actual cases, see K. Pearson, "A Biometric Study of the Red Blood Corpuscles of the Common Tadpole (*Rana temporaria*) from the Measurements of Ernest Warren," *Biometrika*, Vol. 6, pp. 402–419, 1909.

[5] Diagonal cell is the term applied to a compartment of a row extending diagonally across the correlation surface. In symmetrical tables they contain the frequencies for identical values of the subject and relative.

to obtain the products of the frequencies of all of the columns by all the strips, but by doing so a check is obtained for all entries except those of the diagonal cell.

The great advantage of this method is that it replaces mental and pencil drudgery with rapid mechanical calculation. Clipping the movable column by the side of the one with which it is to be compared in the table, one can obtain the products and the sum of the products simultaneously on a Brunsviga,[6] by merely multiplying the successive pairs of frequencies together and allowing the products to accumulate. Of course the frequencies for the diagonal cell can be quickly obtained, by summing the $n(n-1)$ values for the individual column, in precisely the same manner.

Purely as an illustration of method the intra-individual or homotypic correlation for number of seeds developing per pod in a series of Broom plants (*Cytisus scoparius*)[7] collected at Woods Hole in the late summer of 1907 will now be determined.

TABLE I

SEEDS PER POD FOR TWENTY-THREE INDIVIDUAL PLANTS

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	–	–	2	2	1	9	7	10	9	15	9	11	4	3	8	3	3	3	1	–	–	100
2	–	–	—	—	—	—	1	3	3	6	12	6	14	13	12	5	5	11	5	3	1	100
3	–	–	—	—	1	—	2	4	2	4	10	11	7	14	11	12	13	7	2	–	–	100
4	–	–	—	—	—	5	3	7	12	15	14	16	16	7	3	2	—	—	—	–	–	100
5	–	1	—	1	1	4	5	8	10	10	13	12	10	18	4	2	1	—	—	–	–	100
6	–	–	7	7	7	12	8	11	8	8	7	9	7	7	2	—	—	—	—	–	–	100
7	–	–	—	1	1	7	2	5	5	10	7	16	7	9	10	9	2	8	1	–	–	100
8	1	1	2	1	4	3	4	9	8	8	16	9	9	14	7	4	—	—	—	–	–	100
9	1	–	—	—	1	7	9	4	7	11	18	12	10	7	4	6	3	—	—	–	–	100
10	–	–	2	1	3	5	8	9	7	6	14	11	12	10	6	5	1	—	—	–	–	100
11	–	1	2	2	4	3	8	8	8	8	10	13	11	6	5	5	3	3	—	–	–	100
12	–	–	2	1	6	10	17	17	15	10	11	4	6	1	—	—	—	—	—	–	–	100
13	–	–	—	1	—	1	2	3	12	14	12	14	10	12	12	3	3	1	—	–	–	100
14	–	–	—	1	5	5	9	6	11	11	9	9	11	6	9	5	2	1	—	–	–	100
15	–	–	—	—	—	5	3	6	5	4	8	10	8	8	13	13	9	5	3	–	–	100
16	–	–	1	—	1	7	6	7	7	7	9	8	8	8	8	9	8	4	2	–	–	100
17	–	–	—	1	1	8	9	10	7	11	10	14	13	12	3	1	—	—	—	–	–	100
18	–	–	—	—	—	2	1	3	6	12	15	13	13	7	15	8	4	1	—	–	–	100
19	–	–	—	—	1	5	5	7	15	12	11	14	13	11	4	2	—	—	—	–	–	100
20	–	–	—	3	4	6	10	12	6	6	12	10	11	6	6	3	3	2	—	–	–	100
21	–	–	2	2	1	1	4	4	8	5	11	13	10	9	7	3	3	1	—	–	–	84
22	–	1	2	5	5	13	8	8	6	13	5	6	6	2	—	—	—	—	—	–	–	80
23	–	–	2	14	12	11	11	9	6	7	1	1	—	—	—	1	—	—	—	–	–	75
	2	4	24	43	59	129	142	170	183	213	244	242	216	190	149	101	63	47	14	3	1	2,239

[6] A Comptometer will also do.

[7] The variability of these has already been compared with that of Pearson's English series. See "Variation in the Number of Seeds per Pod in the Broom, *Cytisus scoparius*," AMER. NAT., Vol. 43, pp. 350–355, 1909.

TABLE II

INTRA-INDIVIDUAL OR HOMOTYPIC CORRELATION FOR SEEDS PER POD IN THE BROOM

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
0		1	2	1	5	10	13	13	15	19	34	21	19	21	11	10	3					198
1	1		6	9	14	23	25	33	32	39	44	40	36	40	16	11	4	3				376
2	2	6	58	105	122	201	196	232	197	207	212	207	173	147	88	51	28	18	4			2,254
3	1	9	105	256	287	390	374	390	313	373	276	300	240	195	123	74	36	30	3			3,775
4	5	14	122	287	286	477	536	559	482	511	489	446	412	320	208	147	70	48	6			5,425
5	10	23	201	390	477	898	1,009	1,130	1,066	1,268	1,204	1,227	1,049	824	586	424	223	166	45			12,220
6	13	25	196	374	536	1,009	1,066	1,284	1,230	1,324	1,435	1,328	1,220	943	649	458	265	161	39	3	1	13,559
7	13	33	232	390	559	1,130	1,284	1,338	1,446	1,581	1,749	1,677	1,513	1,267	885	598	351	253	70	9	3	16,381
8	15	32	197	313	482	1,066	1,230	1,446	1,516	1,835	1,990	1,965	1,762	1,470	1,065	656	365	240	62	9	3	17,719
9	19	39	207	373	511	1,268	1,324	1,581	1,835	2,008	2,296	2,330	2,038	1,694	1,320	814	457	345	89	18	6	20,572
10	34	44	212	276	489	1,204	1,435	1,749	1,990	2,296	2,648	2,708	2,489	2,189	1,671	1,104	669	462	138	36	12	23,855
11	21	40	207	300	446	1,227	1,328	1,677	1,965	2,330	2,708	2,612	2,448	2,170	1,684	1,128	671	494	125	18	6	23,605
12	19	36	173	240	412	1,049	1,220	1,513	1,762	2,038	2,489	2,448	2,094	1,956	1,488	959	575	442	135	42	14	21,104
13	21	40	147	195	320	824	943	1,267	1,470	1,694	2,189	2,170	1,956	1,812	1,397	928	598	458	145	39	13	18,626
14	11	16	88	123	208	586	649	885	1,065	1,320	1,671	1,684	1,488	1,397	1,228	919	618	480	155	36	12	14,639
15	10	11	51	74	147	424	458	598	656	814	1,104	1,128	959	928	919	640	506	361	118	15	5	9,926
16	3	4	28	36	70	223	265	351	365	457	669	671	575	598	618	506	356	275	99	15	5	6,189
17		3	18	30	48	166	161	253	240	345	462	494	442	458	480	361	275	254	103	33	11	4,637
18			4	3	6	45	39	70	62	89	138	125	135	145	155	118	99	103	30	15	5	1,386
19							3	9	9	18	36	18	42	39	36	15	15	33	15	6	3	297
20							1	3	3	6	12	6	14	13	12	5	5	11	5	3		99
	198	376	2,254	3,775	5,425	12,220	13,559	16,381	17,719	20,572	23,855	23,605	21,104	18,626	14,639	9,926	6,189	4,637	1,386	297	99	216,842

The original data appear seriated by individual plants in Table I. From this we derive the symmetrical intra-individual correlation surface, Table II.

Working by the conventional product moment method, but taking all moments around 0 as suggested elsewhere in these pages,[8] we get:

$$\begin{aligned} \Sigma(x') &= 2{,}179{,}781, & \nu_1' &= 10.0524 = A, \\ \Sigma(x'^2) &= 24{,}578{,}235, & \nu_2' &= 113.346284, \\ \Sigma(x'y') &= 22{,}438{,}814, & \mu_2 &= 12.295679 = \sigma^2, \\ N &= 216{,}842. & \Sigma(x'y')/N &= 103.480018. \end{aligned}$$

The y moments are of course the same as the x, and we have:

$$r = \frac{S(xy)}{N\sigma_x\sigma_y} = \frac{S(x'y')/N - \nu_x'\nu_y'}{\sigma^2} = \frac{S(x'y')/N - A^2}{\mu_2} = .198.$$

Or we may use a short formula for the difference method.[9]

$$r = 1 - \tfrac{1}{2}\frac{\sigma_y^2}{\sigma_x^2} = 1 - .8024 = .198.^{10}$$

Where σ_y is the standard deviation of the difference between pairs.

Where the number of individuals in an array is very small the method presents no very marked advantages, but when the arrays are large it may be very useful and its range of applicability very wide.

For instance, one of the tests of the genotype theory of inheritance is to compare the correlation between parents and offspring with that between the parents co-fraternity and the offspring in a population of self fertilizing or vegetatively

[8] AMER. NAT., Vol. 44, pp. 693–699, 1910.

[9] Harris, J. Arthur, "A Short Method of Calculating the Coefficient of Correlation in the Case of Integral Variates," *Biometrika*, Vol. 7, pp. 215–218, 1909.

[10] On the basis of $N = 2239$, $r = .198 \pm .014$, and we may conclude that the individual plants are slightly but distinctly differentiated in their capacity for maturing seeds. In his English Series of Broom, Pearson (*Phil. Trans. Roy. Soc. Lond.*, A, Vol. 197, p. 335, 1901) found $r = .42$. The difference may easily be attributed to the smallness of the Woods Hole series, some, or possibly all, of the individuals of which may have come from the same parent plant. The case is an illustration of arithmetical method only.

propagating individuals.[11] The correlation surfaces are very easily prepared. Two seriation tables, one for the arrays from which the individual parents were drawn and one for the offspring arrays corresponding to each parental fraternity, are prepared. The first table is cut into strips by columns, passed strip by strip over the offspring seriation table, the frequencies which are in juxtaposition are multiplied together and summed simultaneously, and the resulting totals entered in the proper compartments[12] of a correlation table. This may be called an ascendant-descendant correlation surface. It includes both "parental" and "avuncular" relationships. The "avuncular" relationship is the one sought, and is quickly gotten by subtracting the surface for the relationship between individual parents and their offspring (which will have been already prepared for other purposes) from the ascendant correlation surface just described.

In a forthcoming paper I shall show how various correlations may sometimes be most easily determined from the first two moments for the individual classes or families without the labor of drawing up tables. J. ARTHUR HARRIS.

COLD SPRING HARBOR, N. Y.,
July 7, 1911

ACQUIRED CHARACTERS DEFINED

It is believed that if the term "acquired characters" is carefully defined, and the matter considered in view of that definition, a new light will be cast upon a generally misunderstood subject. The things to be defined are the verb *to acquire,* which means to obtain by effort, and the noun *character,* which means something forming part of an individual. The point of view here involved may be illustrated by the following quotation:

> "Some are born great,
> Some achieve greatness, and
> Some have greatness thrust upon them."

This shows three ways in which an individual obtains greatness. The same three ways apply to the different characters

[11] For an illustration see K. Pearson's analysis of Hanel's data for *Hydra grisea,* "Darwinism, Biometry and Some Recent Biology," *Biometrika,* Vol. 7, pp. 368–385, 1910.

[12] The compartments corresponding to the captions of the two columns dealt with.